Task	Question	Answer	Example thread	Added by
Docker Installation	I've come to the docker installation and there is a screen explaining that before the next step I need to authorize virtualization extension in BIOS. It just says "open BIOS".	my guess is that you are running windows on your laptop. The way Docker works under windows is that it has to install a virtual machine running linux on the side. In order to be able to do this, your computer must be setup to enable this. The BIOS is the part of your computer that controls the initial startup when your computer boots. After switching on your laptop (after a full shutdown, not sleep or so), typically the computer will give you a message: press F1 / Esc / ... to enter setup. If you don't see that message at all, you can also search for the name of the manufacturer (e.g. Dell / HP / etc) on the internet with a term like "enter bios" Within the BIOS settings, there should be a setting regarding virtualization which probably is switched off in your case.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/1/threads/KagTkeDqEeiKtRJz-D_0AA	
Docker Installation	How to install the docker bigdatateam/hdfs-notebook?	Run the test container to check the installation successful: docker run hello-world You can see some Docker deployment logs on the screen and then “Hello from Docker!” message from successfully deployed container.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/1/threads/x2r	
1.1 Demo Spark Task	I cant get a passing grade for the demo spark assignment.	Review the video 'How to submit your first assignment' and pay close attention to the text being typed in the video. It tells you everything you need to know.		tghunt: 2018/09/23
1.1: Demo	I'm trying to solve Week 1's assignment and i can't seem to figure out how to move a file from local system to hadoop via the dockers.	You can use hdfs shell commands to transfer data from local FS in docker and vice versa. See the HDFS shell userguide for more details.		
1.3. Peer review	Can't execute any HDFS command from the container: "/bin/sh: 1: hdfs: not found"	Are you trying to use the Spark Sandbox for writing Hadoop Distributed File system commands? For writing hdfs commands please use the HDFS CLI Playground which can be found here : https://www.coursera.org/learn/big-data-essentials/ungradedLti/SGWv1/hdfs-cli-playground The second solution is to install docker locally. To install docker please follow these instructions : https://www.coursera.org/learn/big-data-essentials/supplement/L7Eea/docker-installation-guide	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/1/threads/LSVVxpmyEeil2BJyIH1Zvg	
3.1: Words Rating	Can anybody advise what a second job is? “RES1 description: Check that the first job has more than 1 reducer.”	You should make two MapReduce jobs The first job’s output is an input for the second job. For the second job you also need to create a corresponding mapper and a reducer. It's also important to have both jobs in the same cell. Otherwise, the variable OUT_DIR won't be accessible.		
3.1: Words Rating	Assignment1: WordsRatingTask, passing in local but failing grader.	cat ${LOGS} >&2. Also read the Hint num.2 (https://www.coursera.org/learn/big-data-essentials/supplement/N6a1Y/hint-to-the-stop-words-programming-assignment) to stop words assignment.		
3.1: Words Rating	I see "Streaming Command Failed!" and incorrect answer.	First I would suggest you do the following in the sandbox to check your python code for bugs: `! python -m pycompile reducer.py` (and similar for mapper) The above can be done in a new cell just below your code, make sure to execute the code first so the file is available. Second, you can check for other issues by passing in a line of text by doing the following: `! echo "this is a test line" | python reducer2.py`	https://bigdata-coursera.slack.com/messages/C9VHJ5XAN/convo/C9VHJ5XAN-1536643906.000100/	
3.2: Stop Words	I need some help with the "stop list"-issue, how many jobs do we have in this task? And the second question now I have the following result and I don't know why?	The log of map-reduce job went into txt file. But the grader still needs it in stderr. So, you need to print your log into stderr. Also, make sure that your parser.py prints the answer into stdout.		
3.2: Stop Words	The code is working fine in the environment but the grader gives the following error: Error: java.lang. RuntimeException: PipeMapRed. waitOutputThreads()	As you can see in the Hadoop Streaming userguide "you will need to use "-file" option to tell the framework to pack your executable files as a part of job submission.". In general you can attach to the job not only the executable files. You can also access them later within your mappers and reducers as if were **located in the same directory**. These files are distributed over working nodes (this mechanism is called DistribetedCache). So in your mappers and reducers you should use the relative paths to these files (e.g. `open("my_file.txt")` not `open("/home/my_dir/my_file.txt")`) because "/home/my_dir/my_file.txt" may not exist on the slave nodes. The "/home/my_dir/my_file.txt" works well in Jupyter because there is a single node in this environment so "/home/my_dir/my_file.txt" exactly exists.	https://bigdata-coursera.slack.com/archives/C6XJUAA2U/p1532736729000042	

Task	Question	Answer	Example thread	Added by
3.2: Stop Words	In the Week 3 assignment, is it absolutely required to do the word count in the log as the notebook suggested? What if I write it in a reducer instead?	Yes, you can put counters in the reducer if you wish.	https://bigdata-coursera.slack.com/messages/C9VHJ5XAN/convo/C9VHJ5XAN-1534729442.000100/	
3.2: Stop Words	I am trying to submit week 6 TF-IDF assignment. I am getting the result 0.000350 which I think is acceptable. I am using only mapper and not a reducer.	Yes, now you can use only 1 job with only map stage for this task.	https://bigdata-coursera.slack.com/messages/C9VHJ5XAN/convo/C9VHJ5XAN-1534115982.000039/	
3.2: Stop Words	I got the right answer for the sample data. But my score is 0. And the grader returns incorrect result.	Python2 division of integers does not happen as expected. That is to say, when you execute print(1/2) in python2 you get an output of 0 since both numbers are interpreted as integers. However, if you execute print(1.0/2.0) you get an output of 0.5 since both inputs are interpreted as floats. To get around the issue in execution of your code, you can do one of two things: add `from __future__ import division` Or instead of casting the word counts as int, cast them as float. More information about the python2 division issue can be found here if you are interested: https://www.python.org/dev/peps/pep-0238/	https://bigdata-coursera.slack.com/messages/C9VHJ5XAN/convo/C9VHJ5XAN-1536806733.000100/	
3.x: MapReduce assignments	Grader shows the error: `yarn_api_client.errors.APIError: Response finished with status: 404`	The code didn't start any MapRecuce Job. So the History server REST API in grader couldn't find the corresponding page in JobTracker and returns 404 error.		
3.x: MapReduce assignments	What does this Hint mean: try to redirect all extra output from jobs to /dev/null because it may contain the output from the framework. I'm trying to submit week 1 assignment.	For stages in your process that may output tracing or other log error information you may want to redirect to to /dev/null so it does not end up in the log or on the screen. You can review the stackexchange article about it: https://unix.stackexchange.com/questions/119648/redirecting-to-dev-null. You can always try to submit your job as is, but if you end up with errors in output that does not pertain to the task then redirecting may apply. Let us know if you have any other questions.	https://bigdata-coursera.slack.com/messages/C9VFRF2NQ/convo/C9VFRF2NQ-1534726474.000200/	
3.x: MapReduce assignments	How many MapReduce jobs can we use in assignments?	You can use at most 2 jobs in each assignment		

6.1: Shortest path	how to select those paths which have start vertex = 12 & end vertex = 34. I am able to get the list of all paths, but unable to filter out only those which start at 12 & end at 34.	not a good practice to compute all possible paths and filter only the ones which start with 12 and end with 3
		
		lo this, you can iterate all the paths and get the min length, however additional optimizations and filters may
		
		for path in all_paths:
		if len(path) < min_length:
		min_path = path
		min_length = len(min_path)
		
		nt to achieve a more optimal solution, bellow there is a pseudo for bfs logic, and you can try to implement it
		
		1. initialize parent = {}
		2. queue = [startNode]
		3. While queue is not empty :
		currentNode = queue.pop()
		Foreach neighbour in currentNode :
		if neighbour is not in parent:
		parent[neighbour] = currentNode
		queue.add(neighbour)
		if neighbour==endNode:
		Terminate the while loop
		----- 	
		path = [endNode]
		4. while(parent[endNode]!=null):
		path.append(parent[endNode])
		
		5. Output : The path in reverse way is our final path
			https://www.coursera.org/learn/big-data-essentials/discussions/weeks/6/threads/PYskCorBEei7CxlAToWC9g?sort=createdAtDesc	

6.1. Shortest path	My shortest path assignment submission is keep failing with error "Container killed by YARN for exceeding memory limits. 1.6 GB of 1.5 GB physical memory used. Consider boosting spark.yarn.executor.memoryOverhead".	Since the connections can be undirected, visited vertexes should be filtered before discovering next connections. This way join records and thus memory can be reduced!	https://bigdata-coursera.slack.com/messages/C9VFRF2NQ/convo/C9VFRF2NQ-1532075031.000186/	
6.1: Shortest path	I have the correct result in the sandbox but 12,14,34 on the grader.	Please make sure that you use the correct dataset path. For this task it should be **/data/twitter/twitter_sample_small.txt**	https://bigdata-coursera.slack.com/messages/C9VFRF2NQ/convo/C9VFRF2NQ-1529502242.000320/	

Task	Question	Answer	Example thread	Added by
6.1: Shortest path	I am getting 'Testing (num. 803): test CRS803_1 failed on line "SyntaxError: invalid syntax"!' when I submit my python2 notebook.	As of 2018/11/10 when using 'import sys print(sys.version_info)', the sandbox environment reports: "sys.version_info(major=2, minor=7, micro=12, releaselevel='final', serial=0)" and the grader reports "sys.version_info(major=3, minor=4, micro=3, releaselevel='final', serial=0)". This forces the need to use python 3 coding styles such as print("..") instead of python2 print ".."		tghunt: 2018/11/10
6.2: TDF-IDF	Different results for local and grader output	The online book works with the stop words file within the dataset folder, while the grader works with the stop_words_en.txt file		
6.2: TDF-IDF	Can't pass RES1_6 test however the number of the reducers is correct	Please make sure that you print Hadoop logs into stderr (cat $LOGS >&2). If no the grader won't be able to read the Hadoop logs. Please make sure that you have Hadoop logs in stderr in the last cell of the mnotebook.		
General	How soon after completing the course are certificates available?	The following two things must be done: https://learner.coursera.help/hc/en-us/articles/209819053-Get-a-Course-Certificate https://learner.coursera.help/hc/en-us/articles/208280196-Course-Certificates If no certificate arrives you may review this: https://learner.coursera.help/hc/en-us/articles/209819063-Solve-problems-with-Course-Certificates From the personal experience it has been 24 hours or less Does anyone know the exact times?	https://bigdata-coursera.slack.com/messages/C6XJUAA2U/convo/C6XJUAA2U-1537542131.000100/	tghunt: 2018/09/23
General	My code works fine but in grader I catch syntax error submitting in the grader	Please rename the notebook (remove parentheses, spaces etc.).	https://bigdata-coursera.slack.com/messages/C9VFRF2NQ/convo/C9VFRF2NQ-1536221700.000100/	
General	My submission hangs forever. How to kill or cancel it?	1. Go to the "task status" page in the grading system. 2. Press "refresh". 3. Click on the trash icon. Anyway the job will be killed automatically in a hour. If you still can't submit the new task after 1 hour, please write a request to teaching staff.		
General	A new sandbox for a course after working in another course. IE: had an open sandbox for course 1 and then tried to open a sandbox for course 2 but is being redirected to the sandbox for course 1.	The issue is related to Coursera's security policies and work flow. An open sandbox gets a unique Id that is required to have a submission on it before a unique Id can be assigned for the other course. The way to clear it is to submit something.		tghunt: 2018/11/10
General	Student reports that they have a NameError: name 'unicode' is not defined" issue	As of 2018/11/10 it appears that python 3 is being used for executing python 2 notebooks. When this occurs, both unicode() and decode at not availble functions. In some cases it might be able to complete the task without either of these, such as course 1, week 5 pairs.	https://bigdata-coursera.slack.com/messages/C9VFRF2NQ/convo/C9VFRF2NQ-1541848835.019400/	tghunt: 2018/11/10
General	How to use online tools to execute examples?	1. For Hadoop tasks you should use : HDFS CLI Playground 2. ▼For executing Spark tasks you should use the Sparj Jupyter Sandbox 3. If you want to deploy the environment to your local machine you can use docker (See instructions)	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/QxhF86JEEeik-hK7GwXb2A	
Spark Assignment. Pairs	RES1 test is failing.	When you're working on the assignments you most likely write the code base on the demo assignment. If yes, please delete execfile(...) cell from the notebook. Also set more than 1 core for Spark application (local[2] instead local).	https://bigdata-coursera.slack.com/messages/C6XJUAA2U/convo/C6XJUAA2U-1543865447.018900/	Tetlanesh: 2018/12/03
General	YARN - need more details understanding	There is an excellent series of Cloudera posts about YARN, it covers architecture, scheduling and configuration http://blog.cloudera.com/blog/2015/09/untangling-apache-hadoop-yarn-part-1/ There is also a good tutorial by Apache Hadoop https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/WritingYarnApplications.html	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/AMNW19OcEeeDgqph8eWrA	
General	Need Unix VM to practice commands	You can play with unix terminal within any of our docker containers or within grading environment. In general, when you're working in Jupyter notebook, you can go to the terminal by clicking "New -> Terminal" as you can see on the screenshot on this thread	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/1/threads/R6a5Z85KEeexuwrR6wbv6A	
General	When I quit docker and reboot docker with 'docker run ' command all of the files I saved in container disappeared. Is there any way to save them in jupyter notebook provided in the class?	I found out that we could use our local directory if we connect to the container with some additional variable . docker run -it -p 8888:8888 -v /c/User/<your id on Computer>/<directory path you want>:/home/jovyan/bigdatateam/yarn-notebook this connection will make you can use <directory path you want> as a starting directory in the docker machine (/home/jovyan/)	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/OF2zYr1EEeeJEwo-j1k8xg	
General	Where is a stop words file in the grader's cluster? IOError: [Errno 2] No such file or directory: '/datasets/stop_words_en.txt'	Do you provide the file to distributed cache? Please, relook week 2, especially https://www.coursera.org/learn/big-data-essentials/lecture/AOZT9/distributed-cache . It is about how to use files from a local system in MapReduce application.	https://bigdata-coursera.slack.com/archives/C9VHJ5XAN/p1569330230022800	

Task	Question	Answer	Example thread	Added by
	In the System Testing we put --config $HADOOP_EMPTY_CONFIG variable during streaming. I have tried it and I could found some additional information came up . However, I couldn't interprete this messages. Why do we put --config $HADOOP_EMPTY_CONFIG when we want to do System Testing?	I have used $HADOOP_EMPTY_CONFIG to execute MapReduce on the local machine by emulating Hadoop environment. You need to find (locate) the empty.conf in the local file system and set HADOOP_EMPTY_CONFIG appropriately.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/Ns-Ujb4aEeekKQ55-ZS8LA	
	In the lecture video, combiner script is basically the same as reducer script. This script is designed to process sorted input data, and that is why shuffle and sort phase is necessary prior to reducer phase. So my question is: is there a hidden sort phase between mapper and combiner?	We will have a mapper per block of data. So if we aggregate the word counts in the the mapper, we will have something like ('a',3), ('b',4) from block-1, and something ('a',1), ('c',1) from block-2. If we have combiner, then the sort/shuffle happens within the memory of the data node, combiner processes the data (within the memory), and finally the the data is written to disk. The combiner will output the result to disk locally (in the respective data node where it runs). Sort shuffle phase happens Finally the sorted data is written to nodes where the reducers will run.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/9ZJfh7kdEee0JRLpTox6jg	
	How could we change inmemory_bigram_reducer.py?	Indeed, you have to change the reducer script to reach memory efficiency. In any case you have to play with command line arguments to make sure that the data you get on the reducer is sorted by second word. But if you don't change the reducer script, then you have to load all this data for each (first) word in memory. By having data sorted by first and second words (by bigram), you can stream data and only count how many times you saw the pair. If you have the same output with and without -D flag specified, then it could probably happened by because of the size of the dataset. When it is small, all the data could go to a few reducers and all the data related to one (first) word is always placed on one reducer.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/hK8JBr7eEeecDRJyap2_xg	
	I cant work on week 2 because I dont know where HADOOP_STREAMING_JAR is located	"/opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.8.1.jar"	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/bO1ZpPa_EeejsA6eWGItZg	
	Where can I find the wikipedia file used in the videos, with the same format ?	You'll find wikipedia dump file with a format article_id <tab> article_text in "Real-World Applications: TF-IDF" assignment of the sixth week.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/UTgpUUrTEeiylAp3CA6UnA	
	Why it is applied in between the mapper and reducer because why don't we just partition the data before it gets sent through the mapper?	When we process the data with a mapper, we get the key/value pairs. The key and the value can be completely different from the input record. So we partition the data by the output of the mapper	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/42j8JYWgEeibBw55Z5R0gg	
	1. If it does keep failing, in a real world application, how do you take note of this failure? How does a dev, once deploying his map-reduce job - keep track of what is happening in the job? 2. I mentioned: "Is there an upper limit to the number of failures the Nodemanager oversees before it kills the job?" Am I correct in my assumption that it is the NodeManager - which oversees this mapper/reducer? Or is it some other actor in the Map-reduce execution?	1. The maximum of attempts of mappers and reducers is configurable option. You can change it using "mapreduce.am.max-attempts" of YARN config. On our grading clusters this option has a value "2". It means that the failing mapper will be executed 2+1 = 3. So if the mapper is always failing the job will be killed when each mapper will fail 3 times. 2. As for killing the job by Nodemanager. The NodeManager works on Node level but the the job kills on the cluster-level (i.e. the cluster kills all containers belong to the job). So the Nodemeneger can't kill the job. The ApplicationMaster do it.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/PZqKsbojEeiXQxL6kY9zUg	
	The message from my submission is as follows: Your score is 0.	Looks like you've filtered out punctuations and junk symbols incorrectly: "0%however 1" Grade system output: ===== Testing (num. 701): STARTING =====	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/uoGmLIwZEeipLg7Shes0eg	
	How can I identify the number of a distributed file available on each node?		https://www.coursera.org/learn/big-data-essentials/discussions/weeks/2/threads/DV	
	While writing data, the NameNode sends a list of DataNodes which forms a pipeline on which the data is written. Is this list for all the blocks of data or for a single block of data?	Every data pipeline is created per block. Why? The data may very well occupy more than the storage available for a single DataNode. Furthermore, the NameNode keeps track of where the blocks are stored. According to the Hadoop documentation, the NameNode retrieves a list of DataNodes using a replication target choosing algorithm. So, changing the list of DataNodes or not is up to that algorithm.	https://www.coursera.org/learn/big-data-essentials/discussions/weeks/1/threads/xjvmIMnoEeiSoRJoAEs0XA	
	!hdfs not found	For writing hdfs commands please use the HDFS CLI Playground which can be found here : https://www.coursera.org/learn/big-data-essentials/ungradedLti/SGWv1/hdfs-cli-playground To install docker please follow these instructions : https://www.coursera.org/learn/big-data-essentials/supplement/L7Eea/docker-installation-guide	https://www.coursera.org/learn/big-data-essentials/discussions/all/threads/LSVVxpmyEeil2BJylH1Zvg	
	I cannot access stop_words file when I am not using docker	Download the file from docker and uploaded it to CLI Playground and you will able to complete the task	https://www.coursera.org/learn/big-data-essentials/discussions/all/threads/L9Ad69PLEei34Qrk-jyo_A	

Task	Question	Answer	Example thread	Added by
	Few tips for faster loading of data from external to managed table	1. Try with different partition properties 2. keep only id, year, month while creating the external table Previously it was taking ages, not it closer to 4 minutes	https://bigdata-coursera.slack.com/archives/C9V18T3AM/p1557594694013100	

Assignment 2: Stop Words	I think I got the right percentage. Also I am using one job only, with 8 reducers. Is there something I am missing here? (counters: Total Words = 11937375, Stop Words = 4966319)	See on the example in the task:
		
		print >> sys.stderr, "reporter:counter:Wiki stats,Total words,%d" % count
		
		Why you wrote?
		
		if word in stopwords:
		print >> sys.stderr, "reporter:counter: Wiki Stats, Stop words,%d" % 1
		print >> sys.stderr, "reporter:counter: Wiki Stats, Total words,%d" % 1
		else:
		print >> sys.stderr, "reporter:counter: Wiki Stats, Total words,%d" % 1
		You added lead spaces but for what?
		if word in stopwords:
		print >> sys.stderr, "reporter:counter:Wiki Stats,Stop words,%d" % 1
		print >> sys.stderr, "reporter:counter:Wiki Stats,Total words,%d" % 1
		else:
		print >> sys.stderr, "reporter:counter:Wiki Stats,Total words,%d" % 1
		
		This replacement solves the task.
		Spaces is harmful things in naming. Too many traps are hidden and you find one.
		You can find this yourself because all tests that were failed are connected not with the final result but the counters thus you should dig into this topic.
			https://bigdata-coursera.slack.com/archives/C9VHJ5XAN/p1558615663040900	